geneticXchange LIFE. SCIENCE. DISCOVERY.

Integration Bioinformatics

Promises and Challenges

Su Chung, Ph.D.
Associate Scientist
UC San Diego Supercomputer Center
and
Chief Scientific Officer
geneticXchange, Inc

DOE Genome Workshop

31 January 2002

geneticXchange LIFE. SCIENCE. DISCOVERY.

Genome-enabled Bioinformatics

- High-throughput technologies generate massive amount of data.

  genome sequencing, microarray gene expression, mass spectroscopy, …

- Growth of data and databases in the public and private domains is ever more rapid.

  genomics, gene expression profiles, proteomics, pharmacogenomics, literature, clinical trials…

- Proliferation of computational tools for data analysis and processing continues.

  modeling and simulation, statistical analysis, sequence analysis and gene finding, clustering algorithm, protein folding and structure prediction, data Mining, visualization…

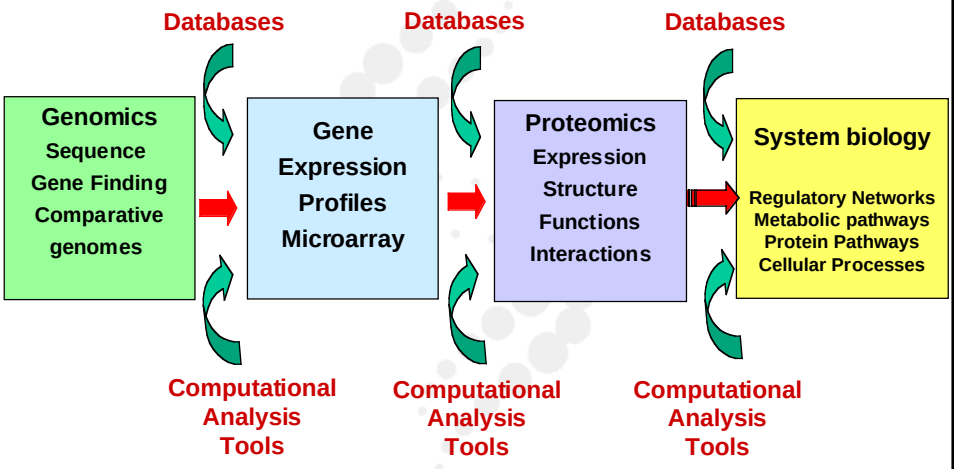
geneticxchange LIFE. SCIENCE. DISCOVERY.

The Driving Force

Data ⇨ Information ⇨ Knowledge ⇨ Discovery

geneticxchange LIFE. SCIENCE. DISCOVERY.

Information-Driven Life Science Research



geneticxchange LIFE. SCIENCE. DISCOVERY.

The Future is Here

- Digitization of biological systems and their processes

  Simulation and modeling of protein-protein interactions, protein pathways, genetic networks, biochemical and cellular processes, normal and disease physiological states,…

- Blurring of the boundary between experimentally generated data and data generated by database searches and computational analyses

- In silico discovery in complement with wet lab experiments

geneticxchange LIFE. SCIENCE. DISCOVERY.

Integration Bioinformatics is becoming the backbone of Research and Discovery

geneticXchange LIFE. SCIENCE. DISCOVERY.

The biological data and databases

- **Complex and Hierarchical**

  Data types range from sequences, 3-dimensional structures, pathways, images, text, and a wide variety of annotation.

- **Heterogeneous**

  storage format, management, and access vary widely

- **Dynamic**

  contents and schema change routinely and rapidly

- **Inconsistent**

  lack standards at the ontology Level

  - Controlled vocabulary for consistent naming for biomedical terms within and between databases
  - Data models for modeling or abstraction of biological system and processes

geneticXchange LIFE. SCIENCE. DISCOVERY.

What is Ontology?

An ontology is a specification of a conceptualization

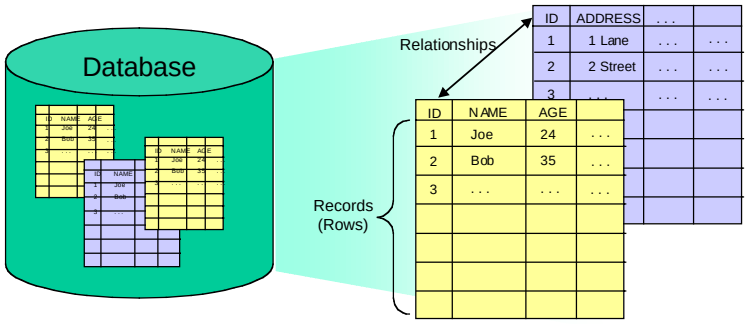
Tom Gruber, Computer Science

Ontology provides a vocabulary for representing and communicating knowledge about some topic and a set of relationships that hold among the terms in that vocabulary

Biologists

Relational Databases

Tables & Relationships



Semi-Structured

XML

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE MOVIE (View Source for full doctype...)>
- <MOVIE TITLE="Waking Ned Devine" YEAR="1998">
  - <CATEGORIES>
      <GENRE>Comedy</GENRE>
    </CATEGORIES>
  - <ACTORS>
    - <ACTOR>
        <FN>Ian</FN>
        <LN>Bannen</LN>
      </ACTOR>
    - <ACTOR>
        <FN>David</FN>
        <LN>Kelly (I)</LN>
      </ACTOR>
    - <ACTOR>
        <FN>Fionnula</FN>
        <LN>Flanagan</LN>
      </ACTOR>
    - <ACTOR>
        <FN>Susan</FN>
        <LN>Lynch</LN>
      </ACTOR>
    - <ACTOR>
        <FN>James</FN>
        <LN>Nesbitt</LN>
      </ACTOR>
    - <ACTOR>
```

- Exchange format
- Nested data structures

geneticxchange LIFE. SCIENCE. DISCOVERY.

Structured Flat Files

ASN.1

```
LOCUS       AF169225_1    413 aa                PRI     14-MAY-2001
DEFINITION  beta-2-adrenergic receptor [Homo sapiens].
ACCESSION   AAD48036
PID         g5714688
VERSION     AAD48036.1  GI:5714688
KEYWORDS
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 413)
AUTHORS   Rupert,J.L., Monsalve,M.V., Devine,D.V. and Hochachka,P.W.
TITLE     Beta2-adrenergic receptor allele frequencies in the Quechua, a high altitude native population
JOURNAL   Ann. Hum. Genet. 64 (2), 135-143 (2000)
MEDLINE   21141798
PUBMED    11246467
FEATURES             Location/Qualifiers
     source          1..413
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="5"
                     /map="5q31-q33"
                     /cell_type="lymphocyte"
                     /tissue_type="blood"
                     /note="isolated from a heterozygous for a known C/T mutation"
     Protein         1..413
                     /product="beta-2-adrenergic receptor"
     CDS             1..413
                     /coded_by="AF169225.1:17..1258"
ORIGIN               1 mgqpgngsaf llapngshap dhdvtqqrde vwvvgmgivm slivlaivfg nvlvitaiak
                    61 ferlqtvtny fitslacadl vmglavvpfg aahilmkmwt fgnfwcefwt sidvlcvtas
                   121 ietlcviavd ryfaitspfk yqslltknka rviilmvwiv sglxsflpiq mhwyrathqe
                   181 aincyaneтc cdfftnqaya iassivsfyv plvimvfvys rvfqeakrql qkidksegrf
                   241 hvqnlsqveq dgrtghglrr sskfclkehk alktlgiimg tftlcwlpff ivnivhviqd
                   301 nlirkevyil lnwigyvnsg fnpliycrsp dfriafqell clrrsslkay gngyssngnt
                   361 geqsgyhveq ekenkllced lpgtedfvgh qgtvpsdnid sqgrncstnd sll
```

Structured

Nested data

Multiple

data types

geneticxchange LIFE. SCIENCE. DISCOVERY.

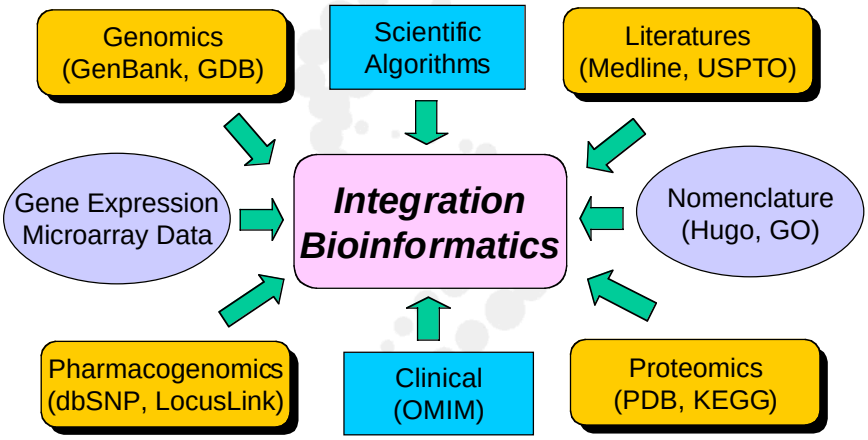
Alias for a Transcription Factor

CEBPB (HUGO Gene Symbol)

- CCAAT/ENHANCER-BINDING PROTEIN, BETA
- C/EBP-BETA
- CRP2
- INTERLEUKIN 6-DEPENDENT DNA-BINDING PROTEIN
- IL6DBP
- NFIL6
- LIVER ACTIVATOR PROTEIN
- LAP
- LIVER-ENRICHED TRANSCRIPTIONAL ACTIVATOR PROTEIN
- TRANSCRIPTION FACTOR 5
- TCF5

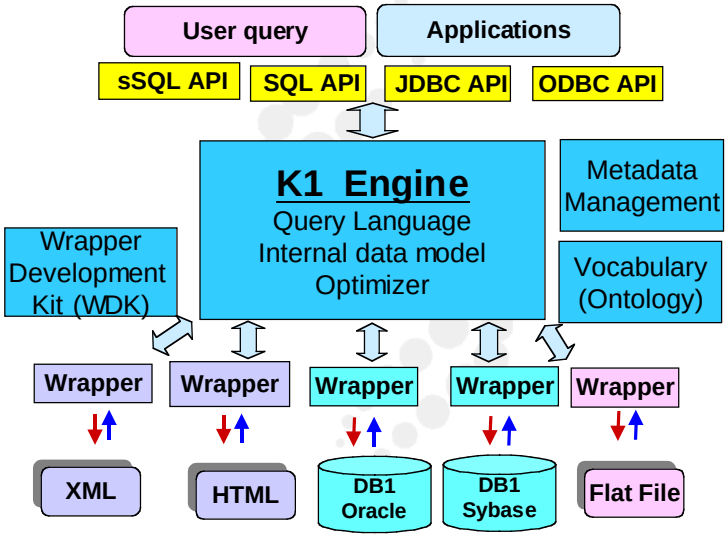
geneticXchange LIFE. SCIENCE. DISCOVERY.

Challenges: Integrate Data Bases and Scientific Algorithms



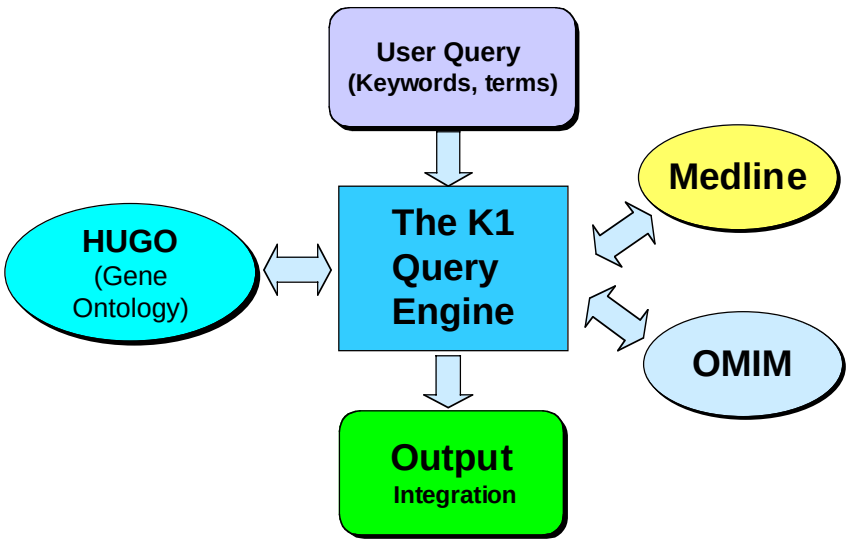
geneticXchange LIFE. SCIENCE. DISCOVERY.

The Mediation Approach: gX Architecture



geneticxchange LIFE. SCIENCE. DISCOVERY.

How does it work? An example



geneticxchange LIFE. SCIENCE. DISCOVERY.

Query Multiple Databases with Ontology

- HUGO (Gene Nomenclature Genew DB)
- OMIM (Online Mendel Ian Inheritance of Man)
- MEDLINE

```
Select
(Hugo: x,
  OMIM:  omim-get-detail (x.MIM),
  PMID1_ABS: ml-get-abstract-by-uid (x.PMID1),
  NUM_Aliases: ml-get-count-general (x.Aliases)))
  from hugo-get-ids() x
  where x.Symbol = "CEBPB";
```

{(#HGNC: "1834", #Symbol: "CEBPB",
#Name: "CCAAT/enhancer binding protein (C/EBP), beta",
#MIM: "189965", #PMID1: "1535333",
#Aliases: "LAP, CRP2, NFIL6, IL6DBP")
#OMIM: {(#uid: 189965,
#gene_map_locus: "20q13.1",….
#allelic_variants: {})},
#PMID1_ABS: {(#muid: 1535333,
#authors: "Szpirer C,…",
#address: "Departement de Biologie…",
#title: "genes encoding the liver-enriched transcription factors C/EBP,…",
#abstract: "By means of somatic cell hybrids segregating either human…..”
#journal: "Genes Dev 1991 Sep;5(9):1538-52")},
#NUM_entries: 1936)}

geneticXchange LIFE. SCIENCE. DISCOVERY.

Advantages

- On Demand Access
  - The most up-to-date, relevant data sources
  - The best-of-breed computational tools
- Real-time information integration for rapid prototyping and decision-making support
- Flexible transformation and manipulation of data
- The right **information** in the right **context**
- In silico discovery

geneticXchange LIFE. SCIENCE. DISCOVERY.

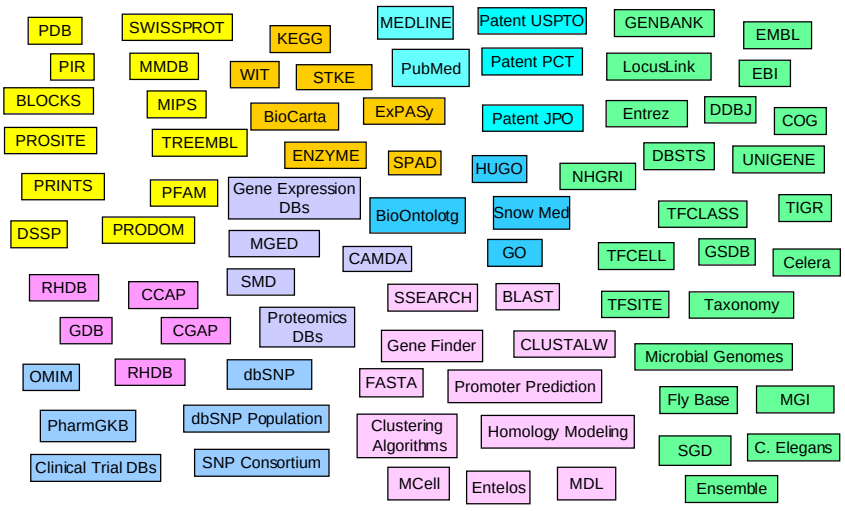
What's in it for the Biologists?

Information Integration

In Silico Discovery

geneticXchange LIFE. SCIENCE. DISCOVERY.

Swimming in Data Sources



geneticxchange LIFE. SCIENCE. DISCOVERY.

What is an In Silico Discovery Kit (ISDK)?

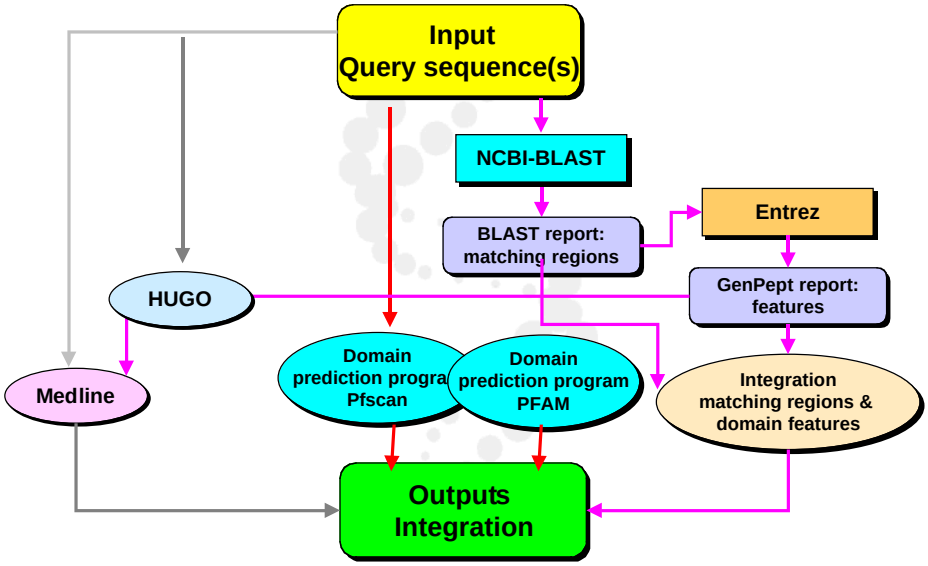
An in silico discovery kit is a script written in the query language that

1. **inputs user data and parameters**
2. **performs a defined information integration task**
3. **output the results**

geneticxchange LIFE. SCIENCE. DISCOVERY.

A Sample of In Silico Discovery Kit (ISDK)

Protein functional domain annotation



geneticXchange LIFE. SCIENCE. DISCOVERY.

ISDKs are the building blocks of in silico discovery

Like Lego-blocks, simple ISDKs can be used to build more sophisticated discovery processes. For example, ISDKs for gene expression analysis, protein functional domain prediction, SNP analysis, and clinical trials can be chained together to form a target identification kit for drug discovery.

geneticXchange LIFE. SCIENCE. DISCOVERY.

ISDKs

- Customizable

  ISDKs provide a base set of templates for bioinformatics integration. These templates can be readily modified and refined to meet user needs to incorporate specific databases and algorithms
- Flexible

  The modular approach of ISDK gives scientists the flexibility to select and combine specific ISDKs for specific research project.
- Scalable: high throughput bioinformatics processing

  The ISDKs are executed automatically by the powerful gX system in batch mode and can handle large data volume
- Reusable codes

  The ISDK scripts are reusable to perform repetitive tasks and can be shared among scientific collaborators
- Updateable

  New databases and new algorithms or computational tools can be readily incorporated into existing ISDK templates

geneticXchange LIFE. SCIENCE. DISCOVERY.

Summary

- A dynamic federated database approach in data integration
- A workflow strategy in information integration
- A plug and play technology that provides non-intrusive enhancement of existing bioinformatics infrastructure
- Innovative in silico discovery kits (ISDKs) that improve the efficiency and productivity of research in the life sciences

geneticXchange LIFE. SCIENCE. DISCOVERY.

contact info:

suchung@sdsc.edu

suchung@ geneticXchange.com